DOCUMENT RESUME

ED 042 816 TM 000 106

AUTHOR	Hambleton, Ronald K.; Traub, Ross E.
TITLE	Analysis of Empirical Data Using Two Logistic Latent Trait Models.
INSTITUTION	Massachusetts Univ., Amherst. School of Education.
REPORT NO	TR-3
PUB DATE	Mar 70
NOTE	34p.; Presented at the annual meeting of American Educational Research Association, Minneapolis, Minnesota, March 1970
EDRS PRICE	EDRS Price MF-$0.25 HC-$1.80
DESCRIPTORS	Achievement Tests, Aptitude Tests, Item Analysis, *Mathematical Models, Scores, *Statistical Analysis, *Statistics, Student Ability, Test Results
IDENTIFIERS	Ontario Scholastic Aptitude Tests, OSAT, SAT, Scholastic Aptitude Test

ABSTRACT

Georg Rasch has developed a new one-parameter latent trait model to explain the performance of examinees on achievement tests. The model can be viewed as a special case of Birnbaum's two-parameter logistic model where all items are assumed to have equal discriminating power. Birnbaum's model permits items to vary in discriminating power. Both models assume that guessing does not occur. This study compares how well the one- and two-parameter models fitted different sets of empirical data. For each model, a measurement of agreement was made between the expected and obtained distributions of ability estimates using three sets of data. The results revealed that the more general the model, the better the fit with real data. (Author/DG)

ED042816

Technical Reports

No. 3

ANALYSIS OF EMPIRICAL DATA USING TWO LOGISTIC LATENT TRAIT MODELS

Ronald K. Hambleton
University of Massachusetts
and
Ross E. Traub
The Ontario Institute for Studies in Education

March, 1970.

CENTER FOR EDUCATIONAL RESEARCH

School of Education
University of Massachusetts
Amherst

TM 000 106

Not to be cited without permission +6

ED042816

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ANALYSIS OF EMPIRICAL DATA USING TWO LOGISTIC LATENT TRAIT MODELS[1,2]

Ronald K. Hambleton

University of Massachusetts

and

Ross E. Traub

The Ontario Institute for Studies in Education

TM 000 106

---

[1]Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minnesota, March, 1970.

[2]The Scholastic Aptitude Test (SAT) data used in the study was kindly provided by Dr. Frederic M. Lord of Educational Testing Service.

ANALYSIS OF EMPIRICAL DATA USING TWO LOGISTIC LATENT TRAIT MODELS

Logistic latent trait models were introduced by Birnbaum (1957, 1958a, 1958b, 1968) for use with binary scored, non-speeded achievement and aptitude tests. For itemized mental tests, a latent trait model specifies a function which relates the probability of success on an item to the underlying latent traits or abilities which the test measures.[1] When a single latent trait is assumed to underlie test performance, the function is usually called an item characteristic curve. The choice of different mathematical forms for the item characteristic curve has led to the development of different latent trait models.

The development of latent trait models rests on two important assumptions. For practical reasons it is usually assumed the items are homogeneous in the sense that they measure the same single ability. According to Lord (1968) this assumption cannot be strictly true for most tests. However it may provide a tolerably good approximation in some instances.

The second assumption is that of "local independence," which implies that the response of an examinee to any test item is statistically independent of his response to any other item in the test. To state it in another way, in an infinite subpopulation of examinees, all of whom are at the same ability level, scores on one test item will be statistically independent of

---

[1]Similar probabilistic models have been used in bio-assay by Finney (1952), Berkson (1953) and others, and in psychophysics beginning with the work of Fechner (1860).

scores on another. [It will be recognized that the assumption of local independence does not imply that test items are uncorrelated over the total group of examinees (Lord and Novick, 1968, p. 361). Correlations between items measuring the same ability will, in general, exist whenever the examinees responding to the items differ on the underlying ability measured by the test.]

Brief Description of the Logistic Models

Two Parameter Logistic Model

Birnbaum proposed a latent trait model in which the item characteristic curves take the form of two-parameter logistic distribution functions,

$$P_g(\theta) = [1 + e^{-Da(\theta-b_g)}]^{-1}, \quad g = 1, 2, \ldots, n.$$

In this equation, $P_g(\theta)$ is the probability that an examinee with ability θ answers item g correctly, a_g and b_g are parameters for item g, $g = 1, 2, \ldots, n$, n is the number of items in the test. The parameter b_g is usually referred to as the index of item difficulty. It represents the point on the ability scale at which the slope of the item characteristic curve is a maximum. The parameter a_g, called item discrimination, is proportional to the slope of $P_g(\theta)$ at the point $\theta = b_g$. The constant D is a scaling factor. [Usually we take $D = 1.7$, to maximize agreement between the logistic model and the normal-ogive model (Lord, 1952).]

Careful inspection of the model reveals an additional implicit assumption characteristic of most latent trait models: guessing does not occur. That this must be so is apparent from the fact that as long as $a_g > 0$, the probability of a correct response to an item decreases to zero as ability decreases.

One-Parameter Logistic Model (Rasch Model)

In the last decade, many researchers have become aware of the work of a Danish Mathematician, Georg Rasch, in the area of latent trait models through his own publications (Rasch, 1960; 1966) and the papers of others advancing his work (Wright, 1967; Wright and Panchapakesan, 1969). Although the Rasch model was developed independently of latent trait theory and along quite different lines, the particular form of the item characteristic curve that he chose can be viewed as a one-parameter logistic model, a special case of Birnbaum's two-parameter logistic model in which all items are assumed to have equal discriminating power and vary only in terms of difficulty. The form of the item characteristic curve can be written as

$$P_g(\theta) = [1 + e^{-D\bar{a}(\theta - b_g)}]^{-1}, \quad g = 1, 2, \ldots, n,$$

in which $\bar{a}$, the only term not previously defined, is the common level of discrimination for the items. The restriction of a common discrimination index results in a set of non-intersecting item characteristic curves which differ only by a translation along the ability scale.

The assumption that all item discrimination parameters are equal is extremely restrictive. Evidence is available which suggests that in at least some tests, unless the items are specially chosen, the assumption will be violated (Birnbaum, 1968, p. 402).

Purpose of the Research

Although Birnbaum's logistic models have been known since 1957, there have been few applications to empirical data reported in the literature. In one study, Ross (1966) found that the two-parameter logistic model fit the data from six tests reasonably well. The tests varied in content, item format,

item difficulty and average inter-item correlations. On the other hand, Wright (1967) reports considerable success in fitting the one-parameter logistic model to test data. However, to this time, no reports have been made of studies comparing the "fit" of the two models to the same set of empirical data.

Other things being equal, it is clear that the two-parameter logistic model should predict the distribution of test scores better than the one-parameter logistic model. This must be so because the two-parameter model makes use of information that the one-parameter model ignores, that is differences among the items in discriminability. Given this fact one might be tempted simply to discard the one-parameter model in favor of the other. However, two valid reasons have been suggested for continuing to use the one-parameter model (see, for example, Panchapakesan, 1969). First, if items can be constructed which satisfy the assumptions of the one-parameter model, then the test score consisting of the total number of correct answers will be a sufficient statistic for estimating the examinee's ability, that is, the number of correct answers contains all the information relevant to the estimation of the examinee's ability. This is not true for more general logistic models (Birnbaum, 1968). Second, there exist fast, numerically efficient procedures for obtaining estimates of item difficulty and ability in the one-parameter model (Wright and Panchapakesan, 1969). Unfortunately, numerically efficient procedures do not exist for estimating the item parameters and abilities in the more general logistic models.

In this study, the two models were compared with respect to their capacity to predict one characteristic of three different sets of test data. The characteristic was the distribution of statistics for estimating ability.

In the one-parameter logistic model or Rasch model the statistic for estimating an examinee's ability is given by the formula, $t = \sum_{g=1}^{n} u_g$, whereas in the two-parameter model the formula is $t = \sum_{g=1}^{n} a_g u_g$. In these formulas, n is the number of items in the test, a_g is the discrimination index of item g and u_g is one if item g has been answered correctly and zero otherwise.

For each set of test data and with each test model the observed and expected distribution of statistics were computed. A measurement of agreement between observed and expected distributions was obtained by using the χ^2 statistic. Comparisons were then made between the χ^2 statistics to determine the gain in prediction by using the two-parameter model rather than the one-parameter model with each set of test data.

Method

Description of Tests and Sample

The three tests chosen for analysis consisted of selected items from the Verbal and Mathematics Sections of the Ontario Scholastic Aptitude Test (OSAT) and the Verbal Section of the Scholastic Aptitude Test (SAT). All three tests were composed of five-option multiple-choice questions. The Verbal Sections of the OSAT and the SAT included antonym, sentence completion and analogy items. In addition the SAT contained reading comprehension items. The Mathematics Section of OSAT included items which called for application of graphical, spatial, algebraic, and numerical reasoning (OSAT Student's Handbook, 1966).

Items in the test which were found to be too easy (i.e. more than 96% of examinees passed the item) or too difficult (i.e. fewer than 4%

of examinees passed the item) were removed. This action can be justified on the grounds that such items provide very little information for the estimation of an examinee's ability (Birnbaum, 1968). Moreover, these items would provide unreliable tetrachoric estimates of correlation with other items. This unreliability might adversely affect the factor analysis of items done subsequently. Only 45 items of the Verbal Section of OSAT and 20 items of the Mathematics Section of OSAT were retained for further analysis. It was not necessary to remove any items from the SAT; however, only the first 80 items in the test were used in the analysis.

For the two sections of OSAT under investigation, a spaced sample (1 in 30) was chosen from the total group of examinees who took the test in Ontario in 1966-67. The resulting sample size was 1319. For the SAT, administered in 1964, a stratified random sample of 1208 examinees was chosen from the sample of 2862 used by Lord (1968). In our sample, the proportion of examinees whose scores fell within score intervals of 10 points (0-0, 10-19, etc.) were the same as the proportions observed in the total group of 103,375 examinees that took the complete test.

Dimensionality of the Tests

One of the assumptions underlying the logistic test models is that the items in the test to which the model is applied measure only a single latent trait. The assumption can rarely be true, but in many practical applications its validity is difficult, if not impossible, to test. This follows from the fact that the dimensionality of a set of items depends, among other things upon the particular choice of correlation coefficient to be analyzed. Analysis of phi correlations, usually leads to more factors than if tetrachoric correlations are used (see the discussion of the difficulty factor problem by McDonald, 1965, 1967). With tetrachoric correlations, the fact that an inter-item inter-correlation matrix has a single factor is a sufficient but,

unfortunately, not a necessary condition for the acceptance of the unidimensionality assumption (Lord and Novick, 1969).

One way to estimate the dimensionality of a set of test items is to perform a principal components analysis on the matrix of tetrachoric item intercorrelations, plot the eigenvalues to estimate the number of common factors, and then perform a principal axis common factor analysis (Green, 1966). The plots of the first 15 eigenvalues for each of the three tetrachoric item correlation matrices are shown in Figures 1, 2, and 3. It is dramatically apparent that the first unrotated factor in each test was dominant. For OSAT-Verbal, the first factor accounted for 22.1% of the total variance, for OSAT-Math, 31.7% and for SAT-Verbal, 20.5%. However, it is certainly true that more than one factor would need to be retained for each test for an acceptable factor solution. For OSAT-Verbal and SAT-Verbal it would probably be necessary to retain at least three factors and for OSAT-Math, at least two factors. Nevertheless, because of the dominant first factor, the decision was made to proceed as if each test was unifactoral.

Item Parameters and Their Estimation

Various methods have been suggested for the estimation of the parameters of the logistic model. The procedure followed here is the one outlined by Lord and Novick (1968). To simplify the computations involved in making the estimates, it is assumed that the underlying ability is normally distributed. It is then possible to express the indices of item difficulty, $\underline{b}_g$, and item discrimination, $\underline{a}_g$, from the two-parameter logistic model in terms of the classical indices of item difficulty, $\underline{\pi}_g$, (where $\underline{\pi}_g$ is the proportion of people in the population correctly answering an item) and item discrimination, $\underline{\rho}_g$, (where $\underline{\rho}_g$ is the biserial correlation between scores on the item

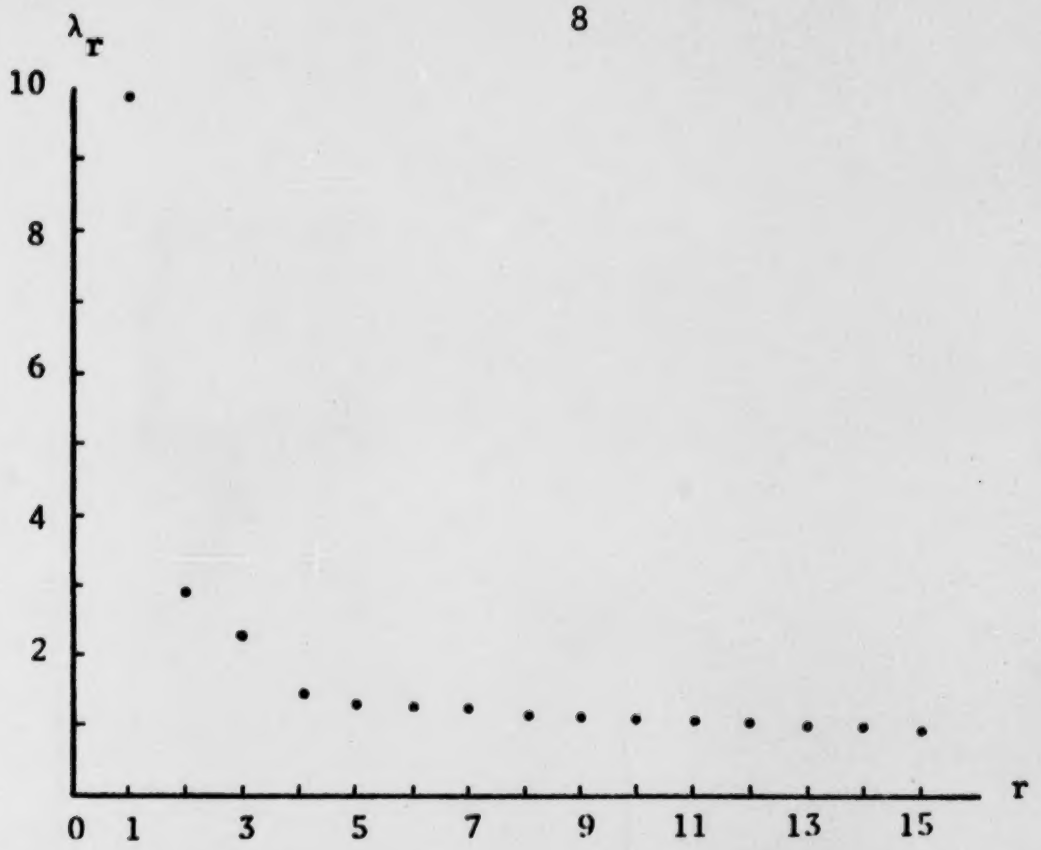

Fig.1--The fifteen largest latent roots λ_r in order of size for the correlation matrix of 45 selected items from the Verbal Section of OSAT.

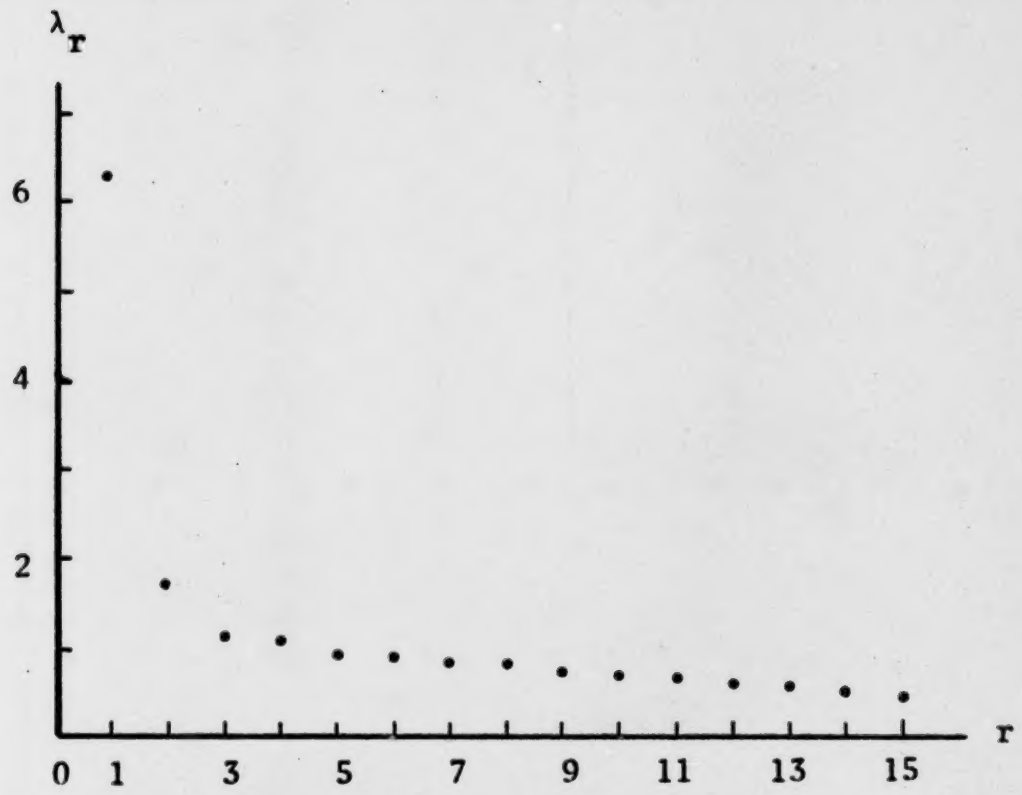

Fig.2--The fifteen largest latent roots λ_r in order of size for the correlation matrix of 20 selected items from the Mathematics Section of OSAT.

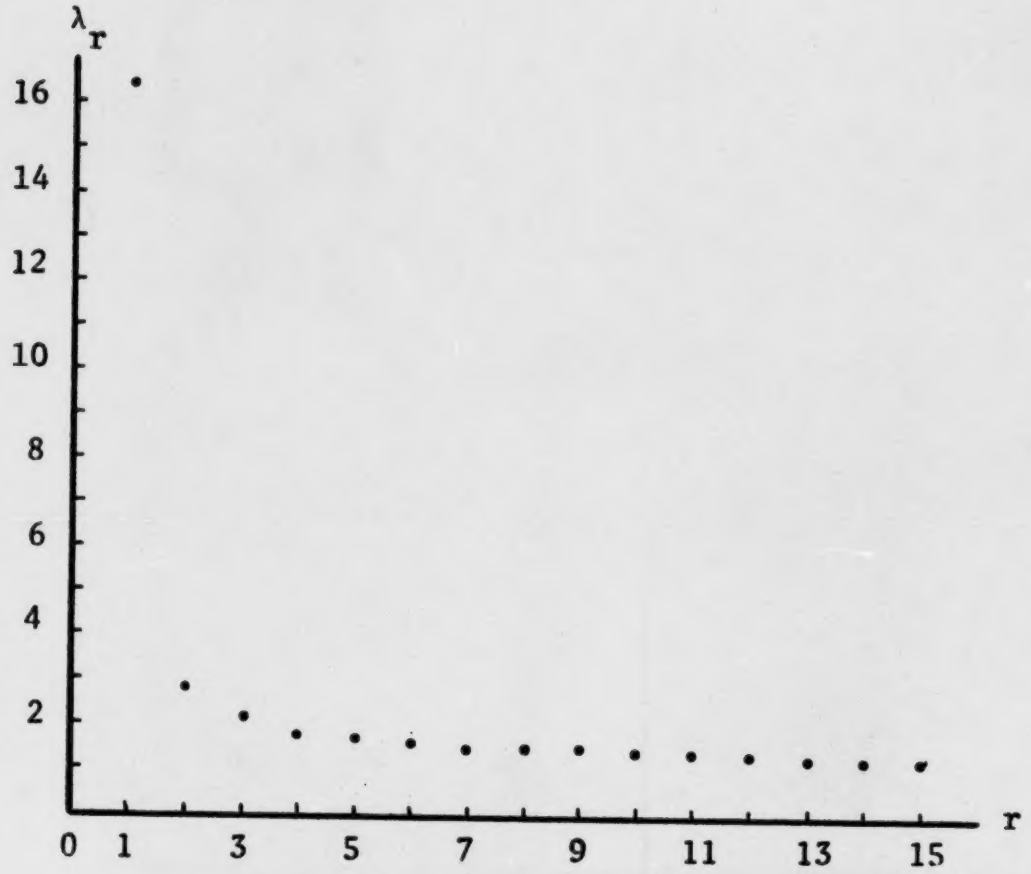

Fig 3--The fifteen largest latent roots λ_r in order of size for the correlation matrix of 80 selected items from the Verbal Section of SAT.

and the latent ability) g = 1, 2, ..., n (Lord and Novick, p. 377). The equations for $\underline{a}_g$ and $\underline{b}_g$ are

$$a_g = \frac{\rho_g}{\sqrt{1 - \rho_g^2}} \quad \text{and} \quad b_g = \frac{1}{a_g} N^{-1}(\pi_g)\sqrt{1 + a_g^2},$$

where $\underline{N}^{-1}(\underline{\pi}_g)$ is the normal deviate corresponding to $\underline{\pi}_g$, g = 1, 2, ..., n.

Assuming that the items measure a single ability, the loading of each item on the first component in the principal component analysis can be taken as an estimate of $\underline{\rho}_g$. This follows from the fact that the loading is the correlation between scores on the item and scores on the first component in the sample of examinees. From $\underline{\rho}_g$, an estimate of $\underline{a}_g$ can be computed. The best estimate of $\underline{\pi}_g$ is the proportion of examinees in the sample who answer item g correctly. With an estimate of $\underline{\pi}_g$ and $\underline{a}_g$, it is then possible to compute an estimate of $\underline{b}_g$. The estimated item parameters $\underline{a}_g$ and $\underline{b}_g$ along with the estimated values of $\underline{\pi}_g$ and $\underline{\rho}_g$ for each of the three tests are summarized in Tables 1, 2, and 3.

In the case of the one-parameter logistic model where it was only necessary to estimate a common level of item discrimination $\bar{a}$ for each item, the geometric mean of the $\underline{a}_g$'s estimated for the two-parameter logistic model was used. (The geometric mean was used rather than the arithmetic mean because it seemed to lead to slightly better predictions of the observed score distributions. However, for none of the three tests, was the difference between the two means greater than .03). In OSAT-Verbal, $\underline{\bar{a}}$ was equal to 0.51; in OSAT-Mathematics $\underline{\bar{a}}$ was equal to 0.66; and in SAT-Verbal, $\underline{\bar{a}}$ was equal to 0.48.

In Table 4 is a summary of certain statistics for each of the three tests under investigation.

TABLE I

Item Parameters Estimated for 45 Selected Items
from the Verbal Section of OSAT
(Sample Size = 1319)

Item	π_g	ρ_g	b_g	a_g	Item	π_g	ρ_g	b_g	a_g
1	.84	.39	-2.57	.42	24	.44	.64	.24	.83
2	.62	.35	- .87	.38	25	.14	.59	1.85	.72
3	.43	.36	.49	.39	26	.14	.63	1.72	.81
4	.19	.32	2.74	.34	27	.32	.46	1.02	.52
5	.19	.37	2.35	.40	28	.32	.52	.95	.61
6	.60	.62	- .41	.79	29	.92	.42	-3.30	.47
7	.53	.58	- .08	.72	30	.66	.43	- .96	.48
8	.63	.50	- .67	.57	31	.40	.37	.67	.40
9	.41	.40	.57	.44	32	.04	.36	4.95	.38
10	.43	.38	.46	.41	33	.23	.49	1.49	.57
11	.70	.65	- .76	.86	34	.27	.46	1.33	.52
12	.49	.61	.04	.76	35	.90	.37	-3.45	.40
13	.77	.45	-1.66	.50	36	.77	.47	-1.51	.53
14	.37	.46	.71	.52	37	.59	.44	- .51	.49
15	.19	.31	2.79	.33	38	.45	.43	.30	.47
16	.15	.32	3.30	.33	39	.27	.53	1.23	.62
17	.81	.66	-1.33	.88	40	.04	.38	4.61	.41
18	.45	.54	.23	.65	41	.81	.46	-1.83	.52
19	.44	.63	.24	.80	42	.62	.34	- .92	.36
20	.62	.41	- .76	.45	43	.39	.50	.55	.58
21	.40	.34	.73	.37	44	.51	.55	- .05	.66
22	.25	.23	3.00	.23	45	.24	.43	1.63	.48
23	.82	.45	-2.06	.50					

TABLE 2

Item Parameters Estimated for 20 Selected Items from the Mathematics Section of OSAT (Sample Size = 1319)

Item	π_g	ρ_g	b_g	a_g
1	.84	.43	-2.34	.47
2	.68	.40	-1.19	.43
3	.82	.52	-1.75	.61
4	.38	.64	.43	.83
5	.20	.58	1.38	.72
6	.77	.36	-2.08	.38
7	.72	.48	-1.30	.54
8	.65	.67	- .57	.91
9	.47	.67	.08	.89
10	.16	.67	1.47	.91
11	.85	.44	-2.27	.49
12	.80	.57	-1.46	.70
13	.70	.54	- .97	.64
14	.60	.74	- .34	1.11
15	.17	.61	1.58	.76
16	.84	.54	-1.84	.64
17	.76	.67	-1.04	.91
18	.47	.56	.13	.67
19	.18	.55	1.65	.66
20	.10	.44	2.78	.49

TABLE 3

Item Parameters Estimated for the First 80 Items
from the Verbal Section of SAT
(Sample Size = 1208)

Item	π_g	ρ_g	b_g	a_g	Item	π_g	ρ_g	b_g	a_g
1	.87	.47	-2.42	.53	41	.50	.42	.00	.46
2	.77	.37	-2.00	.40	42	.60	.29	- .90	.30
3	.78	.50	-1.56	.57	43	.64	.56	- .63	.68
4	.68	.52	- .90	.61	44	.45	.59	.22	.73
5	.45	.40	.33	.43	45	.75	.32	-2.09	.34
6	.41	.57	.40	.70	46	.57	.33	- .54	.35
7	.60	.53	- .48	.62	47	.70	.23	-2.25	.24
8	.55	.23	- .57	.23	48	.83	.41	-2.33	.45
9	.26	.59	1.10	.73	49	.37	.41	.82	.44
10	.27	.63	.98	.81	50	.63	.39	- .86	.42
11	.94	.39	-4.02	.42	51	.83	.49	-1.93	.57
12	.82	.58	-1.58	.71	52	.84	.36	-2.79	.38
13	.75	.63	-1.07	.81	53	.72	.35	-1.68	.37
14	.57	.54	- .32	.65	54	.62	.29	-1.07	.30
15	.28	.43	1.35	.48	55	.64	.28	-1.28	.29
16	.23	.42	1.76	.46	56	.54	.49	- .21	.56
17	.24	.38	1.81	.42	57	.23	.55	1.35	.66
18	.27	.34	1.76	.37	58	.22	.14	5.57	.14
19	.17	.37	2.58	.40	59	.86	.48	-2.24	.55
20	.23	.42	1.76	.46	60	.76	.38	-1.85	.41
21	.75	.37	-2.58	.41	61	.62	.63	- .49	.82
22	.76	.53	-1.76	.63	62	.71	.52	-1.07	.61
23	.66	.45	-1.81	.50	63	.46	.55	.18	.65
24	.76	.39	-1.76	.43	64	.10	.41	3.13	.45
25	.62	.34	-2.58	.36	65	.26	.41	1.58	.45
26	.48	.55	1.76	.66	66	.16	.35	2.86	.37
27	.38	.40	1.78	.44	67	.81	.46	-1.94	.51
28	.30	.53	1.32	.62	68	.80	.45	-1.88	.50
29	.15	.29	.92	.31	69	.76	.41	-1.70	.46
30	.14	.24	1.79	.24	70	.72	.51	-1.13	.60
31	.21	.52	.92	.60	71	.43	.52	.35	.60
32	.30	.52	.09	.61	72	.23	.43	1.71	.48
33	.40	.47	.75	.52	73	.22	.36	2.13	.39
34	.60	.24	-1.00	.24	74	.27	.44	1.39	.49
35	.75	.30	-2.29	.31	75	.25	.60	1.12	.75
36	.27	.39	1.60	.42	76	.68	.39	-1.19	.43
37	.58	.51	- .40	.60	77	.47	.57	.13	.70
38	.21	.49	1.64	.56	78	.50	.44	.00	.49
39	.29	.55	1.00	.66	79	.41	.50	.45	.58
40	.29	.58	.97	.70	80	.38	.50	.61	.57

TABLE 4

Summary Statistics for the Three Tests

Test	Number of Items	Number Correct			Difficulty (b_g)			Discrimination (a_g)		
		Mean	SD	KR-20[1]	Mean	SD	Range	Mean	SD	Range
OSAT-Verbal	45	20.80	7.25	.85	.37	1.81	8.41	.53	.16	.66
OSAT-Math	20	11.16	3.74	.77	-.38	1.47	5.13	.69	.19	.74
SAT-Verbal	80	40.58	11.90	.94	-.02	1.74	9.60	.50	.15	.69

[1]Kuder-Richardson Formula 20 Estimate of Reliability

Expected and Obtained Distributions

A general formula for computing sufficient statistics (called t scores in this study) is $t = \sum_{g=1}^{n} w_g u_g / \sum_{g=1}^{n} w_g$ in which t is a weighted sum of the item responses u_g scored 0 or 1. The w_g's are the weights attached to the items; they depend on the particular choice of a test model. In the one-parameter model the weights are taken to have a value of 1. In the two-parameter logistic model the most information about an examinee's ability is provided if the weights are taken to be Da_g (Birnbaum, 1968).

The task was to predict the distribution of t scores from the best fitting item parameters from the two test models for the three sets of empirical data. The theoretical distributions were derived in the following way. It is seen that

$$E(t|\theta) = \frac{\sum_{g=1}^{n} w_g P_g(\theta)}{\sum_{g=1}^{n} w_g} \qquad (1)$$

and

$$Var(t|\theta) = \frac{\sum_{g=1}^{n} w_g^2 P_g(\theta) Q_g(\theta)}{\left[\sum_{g=1}^{n} w_g\right]^2} \qquad (2)$$

$E(t|\theta)$ is simply the expected t score for an examinee of ability θ. $Var(t|\theta)$ is the variance of t scores for an examinee of ability θ. The expressions in the denominators of (1) and (2) are introduced as scaling factors. In this form, $0 \leq E(t|\theta) \leq 1$. It is clear that $E(t|\theta)$ and $Var(t|\theta)$ depend on the test model since for different models, $P_g(\theta)$ and w_g will differ. The asymptotic distribution (as n, the number of items in the test increases) of t for given θ is normal.

In order to obtain the expected distribution, it was assumed $\underline{\theta}$ was distributed normally, with zero mean and unit variance. This normal distribution was sectioned into 13 parts with boundaries at -3.25, -2.75, ..., 2.75 and 3.25, and the probabilities in each section were assigned to the point -3.0, -2.5, ..., 2.5, 3.0. The values of $E(t|\theta_i)$ and $SD(t|\theta_i)$, where $SD(t|\theta_i) = \sqrt{Var(t|\theta_i)}$, at the points $\theta_1 = -3.0$, $\theta_2 = -2.5$, ..., $\theta_{13} = 3.0$, are summarized in Tables 5, 6, and 7 for the two test models and the three sets of test data.

Since each conditional distribution of $\underline{t}$ given $\underline{\theta}$ is approximately normal according to the theory (provided the test is long enough), the conditional proportion lying between any two points $\underline{t}_i$ and $\underline{t}_j (t_i < t_j)$ on the $\underline{t}$ score scale can be computed. By multiplying the conditional proportion by the probability associated with the corresponding $\underline{\theta}$ and summing across $\underline{\theta}$, the expected proportion of examinees in the sample lying between $\underline{t}_i$ and $\underline{t}_j$ can be computed. If the expected proportion is multiplied by the sample size, the resulting number is the expected frequency of examinees scoring between $\underline{t}_i$ and $\underline{t}_j$.

The observed $\underline{t}$ score of each examinee from the three sets of test data for each of the test models was computed using the formula $t = \sum_{g=1}^{n} w_g u_g / \sum_{g=1}^{n} w_g$. With each model, the appropriate scoring weights were used.

Scores on the $\underline{t}$ score scale were divided into 21 categories. For analysis of OSAT-Math and SAT-Verbal test data, the width of each category was .05. With OSAT-Verbal, the width of each category was 2/45. By correctly locating the lower limit, $\underline{t}_1$, of the bottom category for the analysis of each set of test data it was possible (with minor exceptions) to ensure that the same number of possible test scores, 0, 1, ..., n, fell in each category.

TABLE 5

Distribution of Expected t Scores and Standard Deviation of t Scores for Various Ability Levels with the One- and Two-Parameter Models for OSAT-Verbal

Two-Parameter Logistic Model			One-Parameter Logistic Model		
Ability	$E(t\|\theta)$	$SD(t\|\theta)$	Ability	$E(t\|\theta)$	$SD(t\|\theta)$
-3.0	.09	.04	-3.0	.10	.04
-2.5	.12	.04	-2.5	.14	.05
-2.0	.17	.05	-2.0	.19	.05
-1.5	.23	.06	-1.5	.24	.05
-1.0	.30	.06	-1.0	.31	.06
-0.5	.38	.07	-0.5	.38	.06
0.0	.47	.07	0.0	.45	.06
0.5	.56	.07	0.5	.53	.06
1.0	.65	.06	1.0	.60	.06
1.5	.72	.06	1.5	.67	.06
2.0	.78	.05	2.0	.74	.06
2.5	.84	.05	2.5	.79	.05
3.0	.88	.04	3.0	.84	.05

TABLE 6

Distribution of Expected t Scores and Standard Deviation of t Scores for Various Ability Levels with the One- and Two-Parameter Models for OSAT-Math

Two-Parameter Logistic Model			One-Parameter Logistic Model		
Ability	$E(t\|\theta)$	$SD(t\|\theta)$	Ability	$E(t\|\theta)$	$SD(t\|\theta)$
-3.0	.09	.05	-3.0	.11	.06
-2.5	.13	.06	-2.5	.16	.08
-2.0	.18	.07	-2.0	.23	.08
-1.5	.25	.08	-1.5	.32	.09
-1.0	.34	.09	-1.0	.41	.09
-0.5	.44	.10	-0.5	.50	.09
0.0	.55	.09	0.0	.59	.09
0.5	.65	.09	0.5	.67	.08
1.0	.74	.08	1.0	.74	.08
1.5	.81	.08	1.5	.80	.07
2.0	.87	.07	2.0	.86	.07
2.5	.91	.06	2.5	.90	.06
3.0	.95	.05	3.0	.93	.05

TABLE 7

Distribution of Expected t Scores and Standard Deviation of t Scores for Various Ability Levels with the One- and Two-Parameter Models for SAT-Verbal

Two-Parameter Logistic Model			One-Parameter Logistic Model		
Ability	$E(t\|\theta)$	$SD(t\|\theta)$	Ability	$E(t\|\theta)$	$SD(t\|\theta)$
-3.0	.12	.03	-3.0	.14	.04
-2.5	.16	.04	-2.5	.18	.04
-2.0	.21	.04	-2.0	.24	.04
-1.5	.27	.04	-1.5	.30	.04
-1.0	.34	.05	-1.0	.37	.05
-0.5	.42	.05	-0.5	.44	.05
0.0	.50	.05	0.0	.51	.05
0.5	.59	.05	0.5	.59	.05
1.0	.67	.05	1.0	.66	.05
1.5	.75	.04	1.5	.72	.04
2.0	.81	.04	2.0	.78	.04
2.5	.86	.04	2.5	.82	.04
3.0	.90	.03	3.0	.87	.03

With OSAT-Math, one test score fell in each category, with OSAT-Verbal, two test scores fell in each category with the exception of the extreme categories where four test scores fell in each, and with SAT-Verbal four test scores fell in each category with the exception of the bottom category which contained two test scores and the top category which contained three.

The expected and obtained t score distributions (in the 21 categories) with the one- and two-parameter logistic models were computed for the three sets of test data.

Goodness-of-Fit

As a measure of the goodness-of-fit between the observed and expected score distributions using each test model, with the three sets of test data, we computed the χ^2 statistic,

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i} \qquad (3)$$

where O_i is the number of examinees in the i^{th} score group category, E_i is the expected or predicted number of examinees in the i^{th} score group category under the assumption that the test model under investigation is the true one, and N is the number of categories over which the comparison is to be made. Lazarsfeld and Henry (1968, p. 77) emphasize a point concerning the χ^2 statistic which bears on the present application. The E_i's in equation (3) are not calculated from the true parameters but rather from some estimate of the true parameters. Therefore, it is not suggested that the resulting statistics should be compared with tabled values of the χ^2 distribution. Rather, for a fixed number of categories, N, it is possible to compare the relative size of χ^2 for fitting a given set of data with different models. Applying the rule suggested by

Cochran (1954) categories at the extreme ends of the expected distribution were combined to ensure that the expected numbers of examinees in each category exceeded one. With each set of test data, the number of categories was reduced from 21 to 17 by combining the three categories at each end of the expected and observed score distributions.

Results and Discussion

The goodness-of-fit results between the expected and obtained distributions with the different test models expressed in terms of the χ^2 statistic for OSAT-Verbal, OSAT-Math and SAT-Verbal test data are summarized in Table 8. The expected and obtained distributions with each test model and each set of test data are shown in Figures 4 to 9.

TABLE 8

Goodness-of-Fit Results

Test Data	Two-Parameter Logistic Model χ^2	One-Parameter Logistic Model χ^2
OSAT-Verbal	27.01	35.47
OSAT-Math	14.08	28.82
SAT-Verbal	63.97	70.38

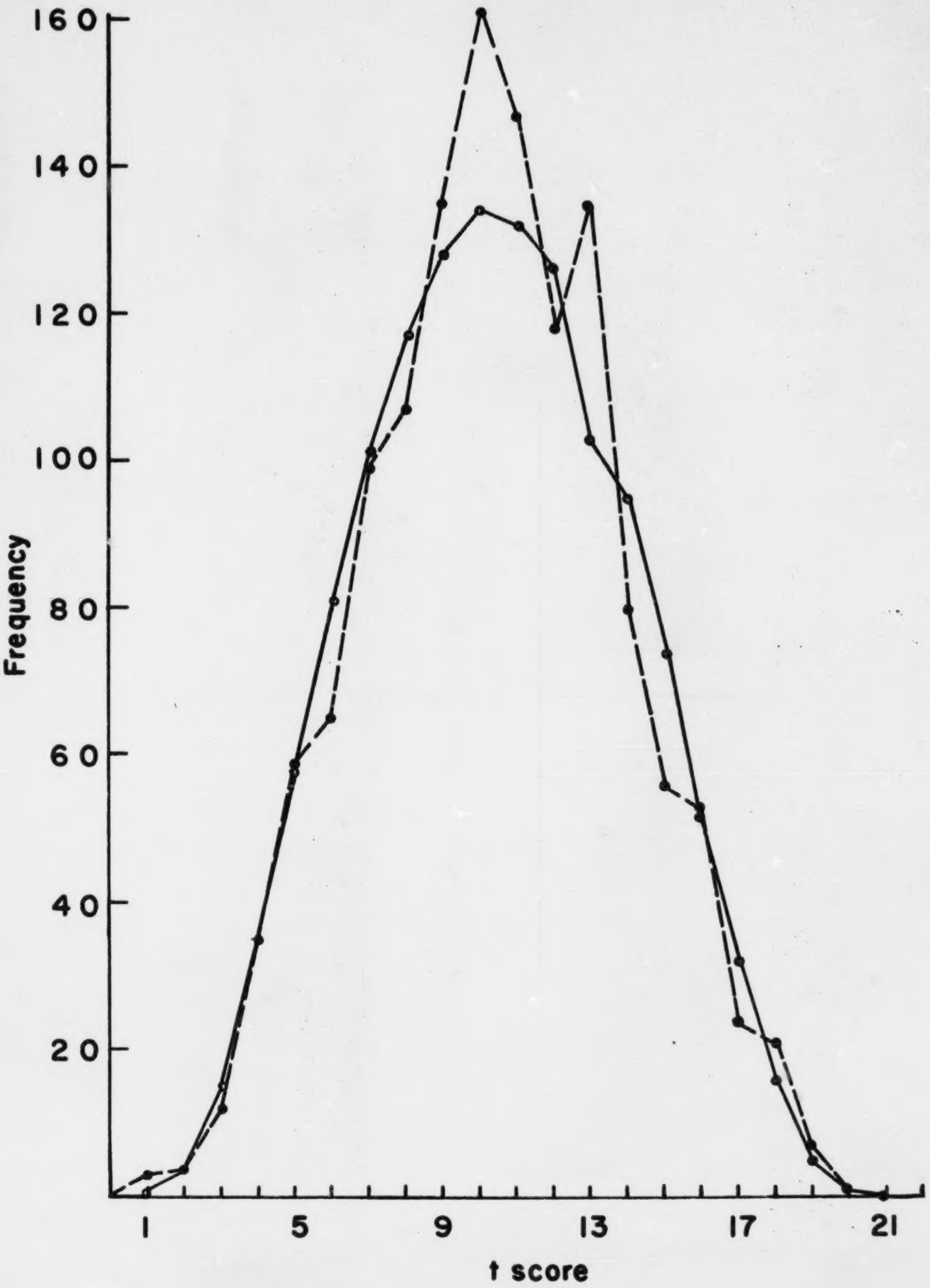

Fig. 4--Observed (--●--) and expected (—○—) distributions for OSAT-Verbal using the two-parameter logistic model.

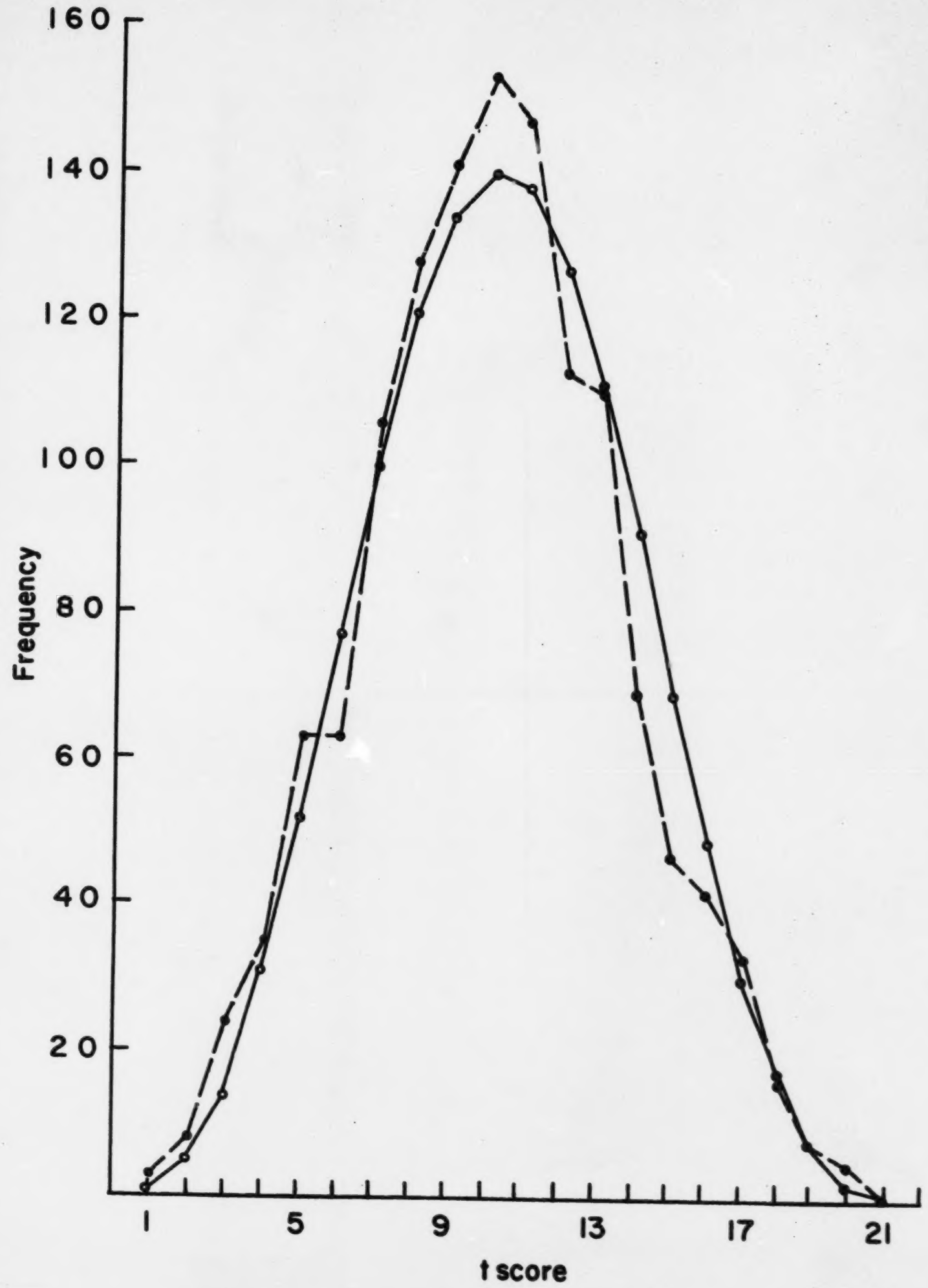

Fig. 5--Observed (--●--) and expected (--O--) distributions for OSAT-Verbal using the one-parameter logistic model.

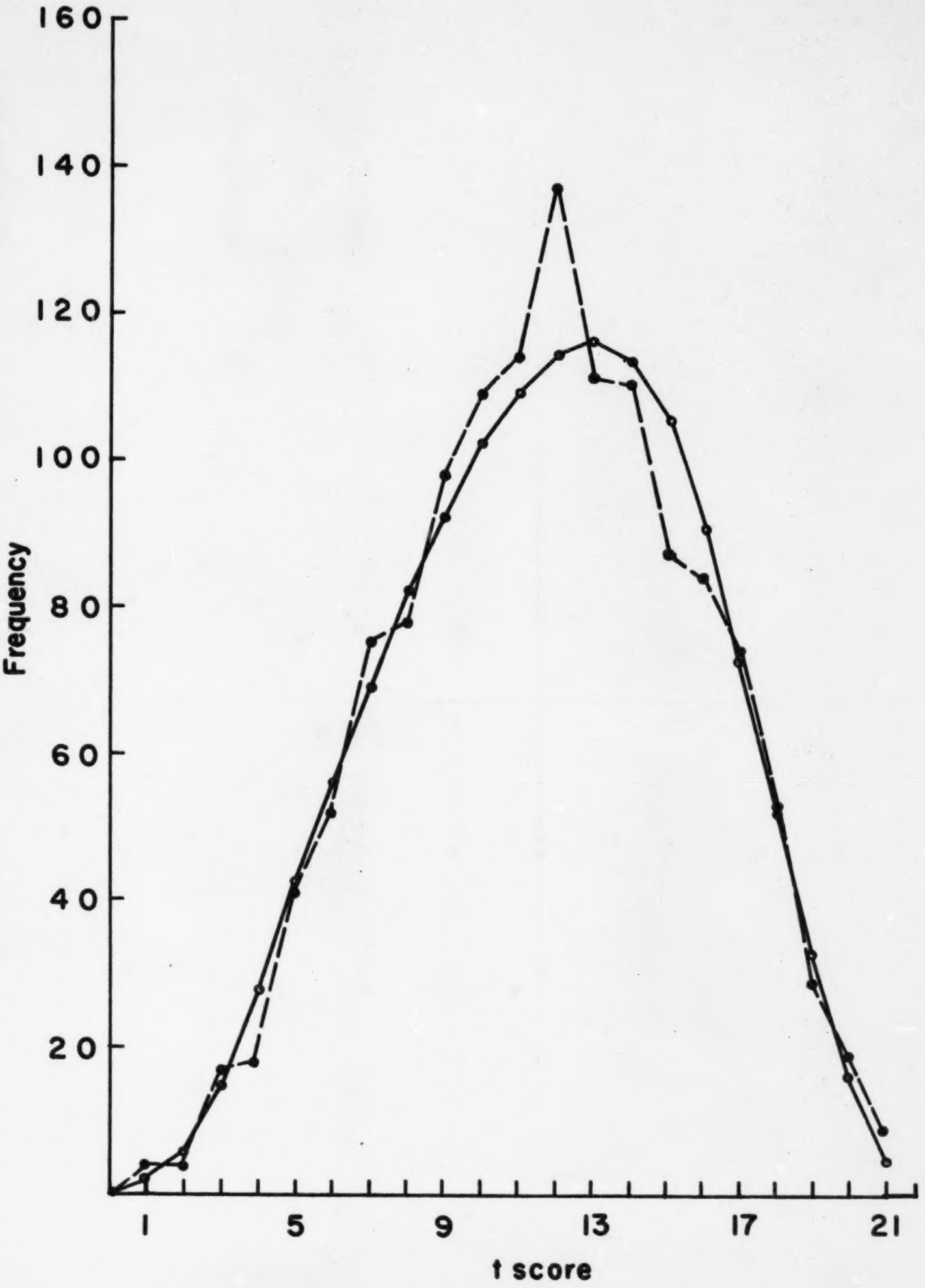

Fig. 6--Observed (--●--) and expected (—0—) distributions for OSAT-Mathematics using the two-parameter logistic model.

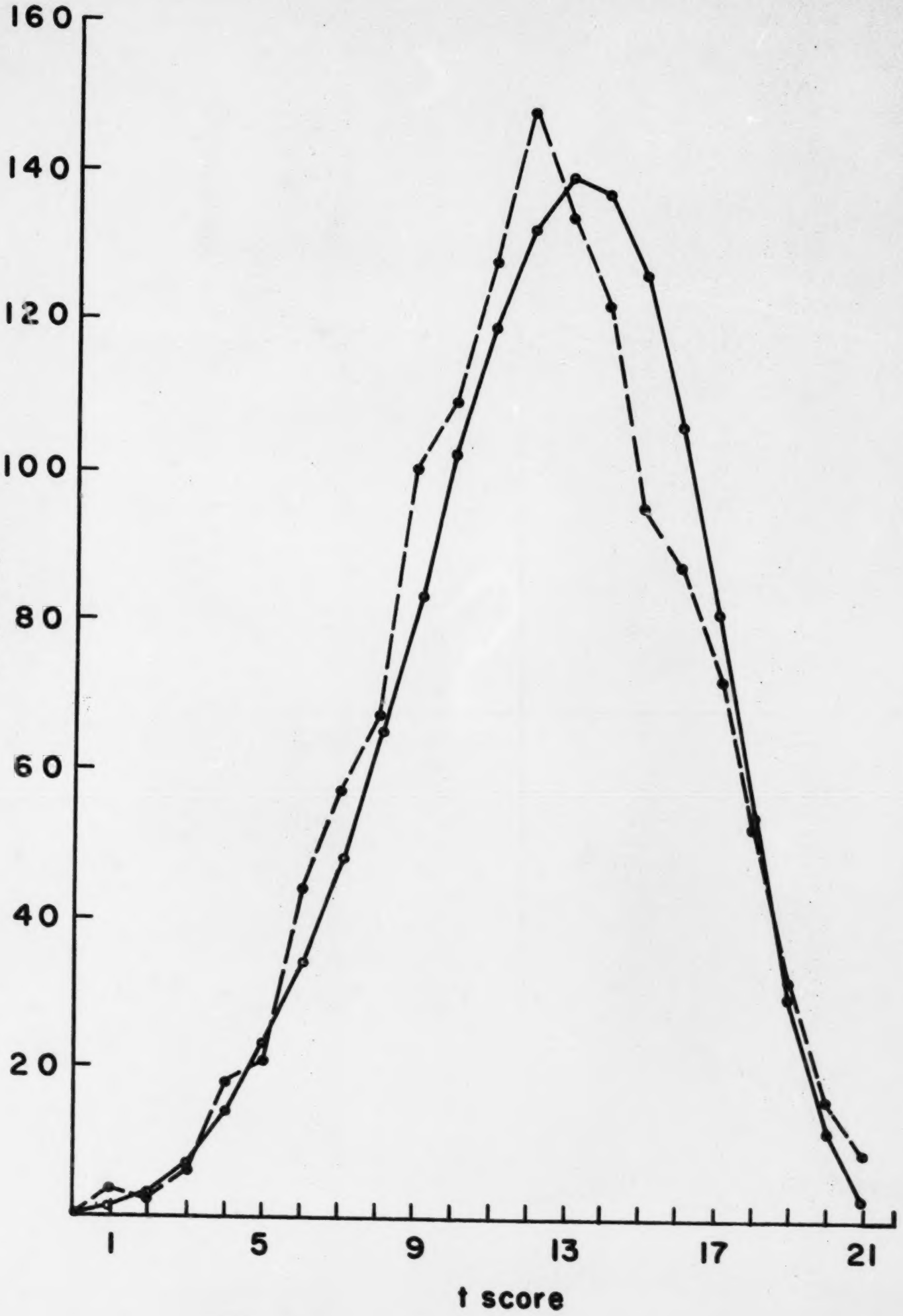

Fig. 7--Observed (--●--) and expected (—○—) distributions for OSAT-Mathematics using the one-parameter logistic model.

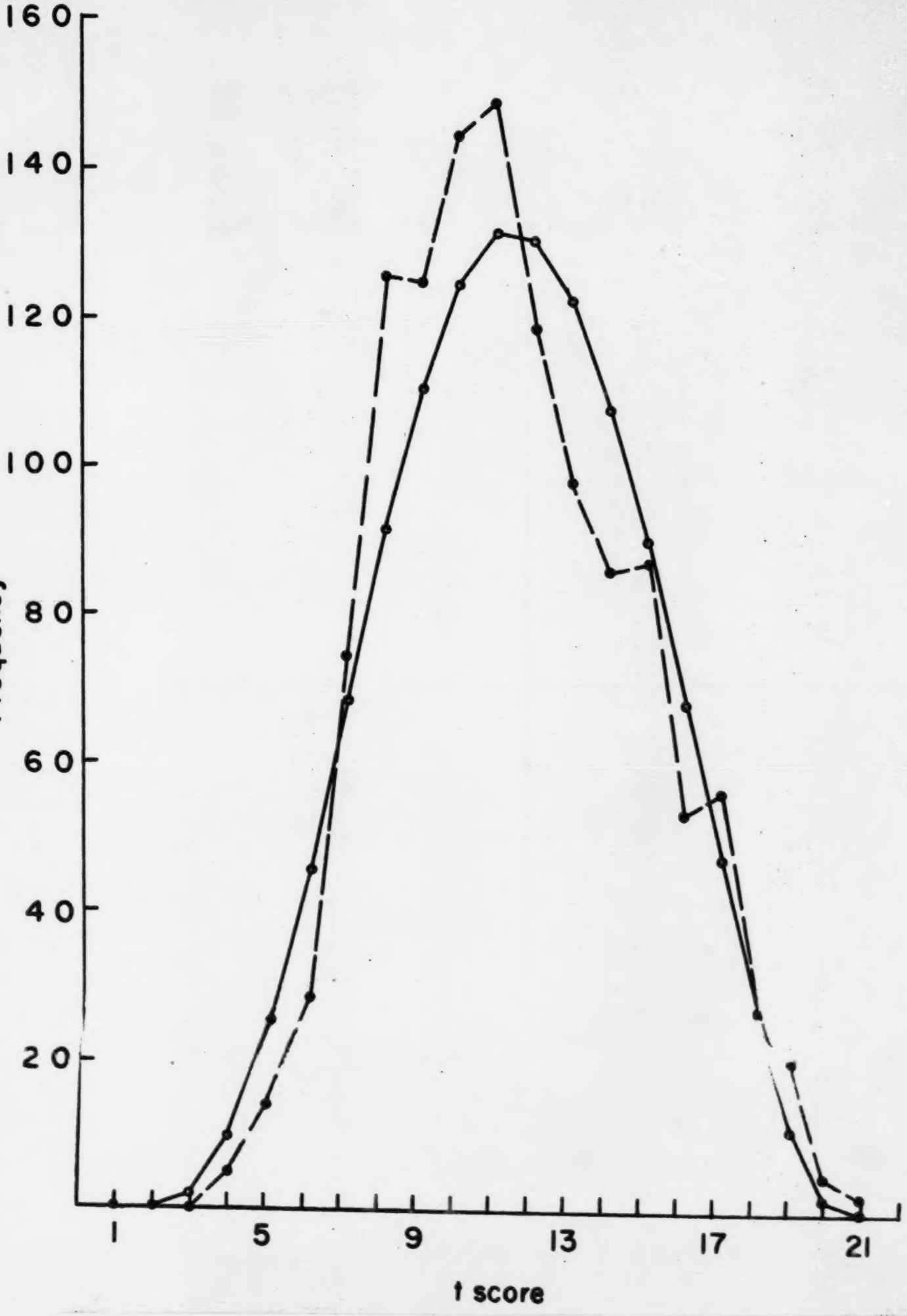

Fig. 8--Observed (--0--) and expected (—0—) distributions for SAT-Verbal using the two-parameter logistic model.

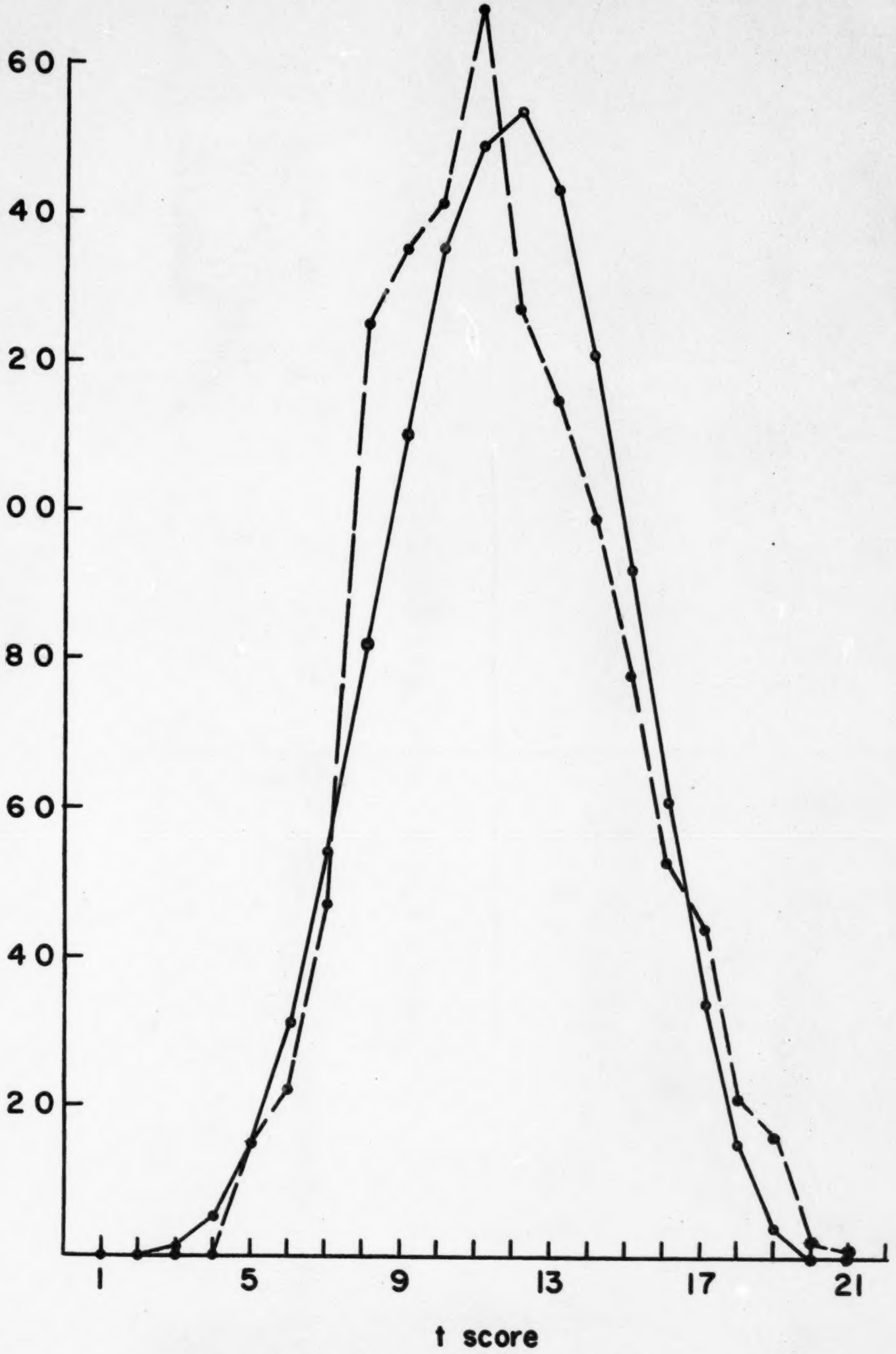

Fig. 9--Observed (--●--) and expected (--0--) distributions for SAT-Verbal using the one-parameter logistic model.

It is apparent from Table 8 that the more general the test model, the better the fit between the observed and expected distributions, regardless of the test. With each set of test data, there are losses in predicting test performance when the one-parameter (or Rasch) model is substituted for the two-parameter model. The loss was greatest on the shortest test (OSAT-Math) and smallest on the longest test (SAT-Verbal). These findings lend support to Birnbaum's conjecture (1968, p. 492) that if the number of items in a test is very large the inferences that can be made about an examinee's ability will be very similar for the one- and two-parameter logistic models.

The goodness-of-fit results reported in Table 8 reveal that generally the best fits were obtained with OSAT-Math, somewhat poorer fits with OSAT-Verbal, and the poorest fits with SAT-Verbal. It is noteworthy that the best fits were obtained with the OSAT-Math test which, as indicated in the methods section, appeared to be the most homogeneous test, i. e. the test which came closest to satisfying the unidimensionality assumption which underlies both logistic test models. The poor fits with the SAT-Verbal test can be explained, partially at least, by an unfortunate choice of samples. The available SAT data was based on the first 80 items of the test, but the observed-score distribution available for the total group that took the SAT was based on the full test. Thus, before it was possible to choose our sample so as to represent the total group, it was necessary to estimate the observed-score distribution of the total group on the shorter test. This was done by scaling down the test scores on the full test by a constant factor. But since the majority of items omitted from the full test were difficult ones because they came at the end of the test, it is apparent that scaling scores down in the way we did would have the effect of producing an estimated score distribution on the 80-item test for the total group that would be

somewhat lower than it should have been, particularly for students of low and average ability. Although this explanation may account, in part, for the poor fits observed between the models and the data, the comparison between the two models in predicting the test scores is still legitimate because the error in estimation affects both test models.

There exist at least two additional extraneous reasons for discrepancies between the observed and expected score distributions under both test models. First, the models apply in theory only to non-speeded tests (Lord and Novick, 1968). None of the three tests used in this study is highly speeded, but the item analyses of the OSAT-Verbal and OSAT-Math tests suggest that they are at least partially speeded. Second, the two-parameter model of Birnbaum (1968) and the one-parameter model of Rasch (1960; 1966) were developed for tests in which guessing has a negligible effect on performance. It is hard to imagine a multiple-choice test where guessing does not play at least a small part in determining the outcome of test scores, particularly for low ability examinees. Birnbaum (1968) suggested an improvement on the two-parameter logistic model by adding a third parameter to take into account the level of guessing on each item. While the three-parameter model has more intuitive appeal than the two-parameter model, it raises the problem of estimating yet another item parameter. Nevertheless, it is highly recommended that in future research the three-parameter model be used for comparative purposes with the one- and two-parameter models in situations where guessing is known to be a factor in test performance.

The foregoing points notwithstanding, the fact still remains that for each set of test data studied, there was a substantial improvement in the agreement between model and data as a direct result of adding an item discrimination parameter to the model. The results of this study suggest that

the two-parameter model will provide greatest improvements over the one-parameter model when applied to data from short tests where the variability of the discrimination parameters is substantial. Whether the gains are worth the increased cost of solving for the parameters of the more complex model is a question which requires investigation.

References

Berkson, J. A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. Journal of the American Statistical Association, 1953, 48, 565-599.

Birnbaum, A. Efficient design and use of tests of a mental ability for various decision-making problems. Series Report No. 58-16, Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, 1957.

Birnbaum, A. On the estimation of mental ability. Series Report No. 15, Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (a)

Birnbaum, A. Further considerations of efficiency in tests of a mental ability. Technical Report No. 17, Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (b)

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores, Reading, Mass.: Addison-Wesley, 1968, Chapters 17-20.

Cochran, W. G. Some methods for strengthening the common χ^2 tests. Biometrics, 1954, 10, 417-451.

Fechner, G. T. Elemente der Psychophysik. Leipzig: Breitkopf and Hartel, 1860.

Finney, D. J. Probit Analysis. London: Cambridge University Press, 1952.

Green, B. F., Jr. The computer revolution in psychometrics. Psychometrika, 1966, 31, 437-445.

Lazarsfeld, P. F. and Henry, M. E. Latent Structure Analysis. Boston: Houghton-Mifflin, 1968.

Lord, F. M. A theory of test scores. Psychometric Monograph, No. 7, 1952.

Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

McDonald, R. P. Difficulty factors and nonlinear factor analysis. British Journal of Mathematical and Statistical Psychology, 1965, 18, 11-23.

McDonald, R. P. Nonlinear factor analysis. Psychometric Monograph, No. 15, 1967.

Panchapakesan, Nargis. The simple logistic model and mental measurement. Unpublished doctoral dissertation, University of Chicago, 1969.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedogogishe Institut, 1960.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Ross, J. An empirical study of a logistic mental test model. Psychometrika, 1966, 31, 325-340.

The Ontario Institute for Studies in Education. OSAT Student's Handbook, 1966.

Wright, B. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1967.

Wright, B. and Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.